Proviso with Pseudowords:
A Between-Subj Version

2020 Nov 24

Link:
https://spellout.net/ibexexps/sherrychen/proviso2020_pilot3_{n,i,d}/experiment.html
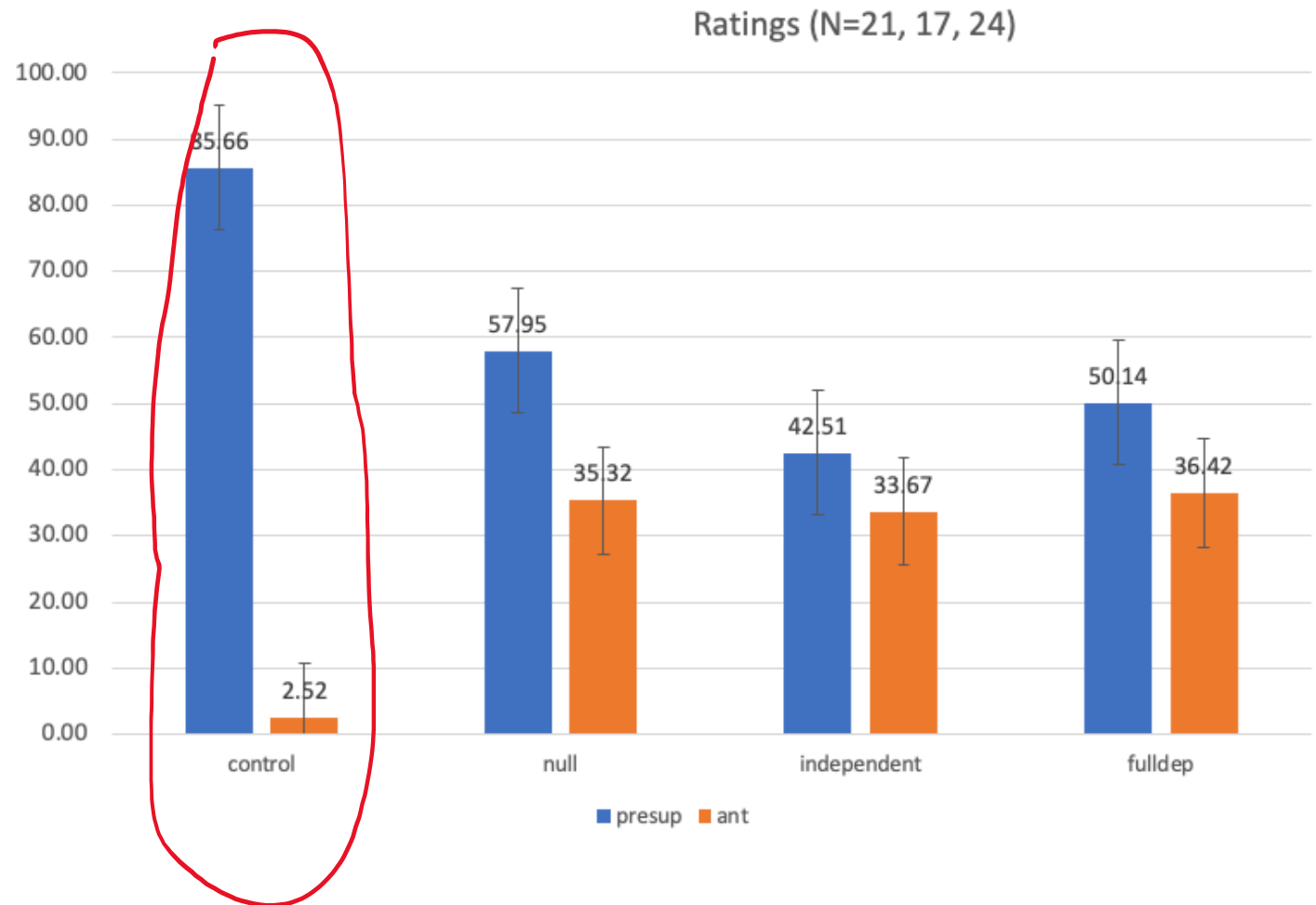
Key Changes from Pilot 2 (Nov 20)

- Context: now **between-subject**
  - Because we suspected an order effect in pilot 2: there was a steady increase in ratings for the ant condition over trials, which likely happened as well in the presup condition as well; ppl possibly started to strategize very quickly after seeing null + indep
  - Current: Each subj sees 2 items only, the control + a critical condition from {null, independent, fulldep}
- **Changed "unexpected" back to "surprised"**
- Simplified narrative: journalistic adventure to 2 planets

Participants

- 119 participants recruited from MTurk (40 for each)
  - 3 non-native speakers: 1606236714, 1606238577, 1606238134
  - Filtering based on the control item:
    - Control.Presup > 50 (70%) **AND** Control.Ant < 20 (70%)
- Remaining: **62**
  - Similar rate as last time (~50%), super noisy!!!
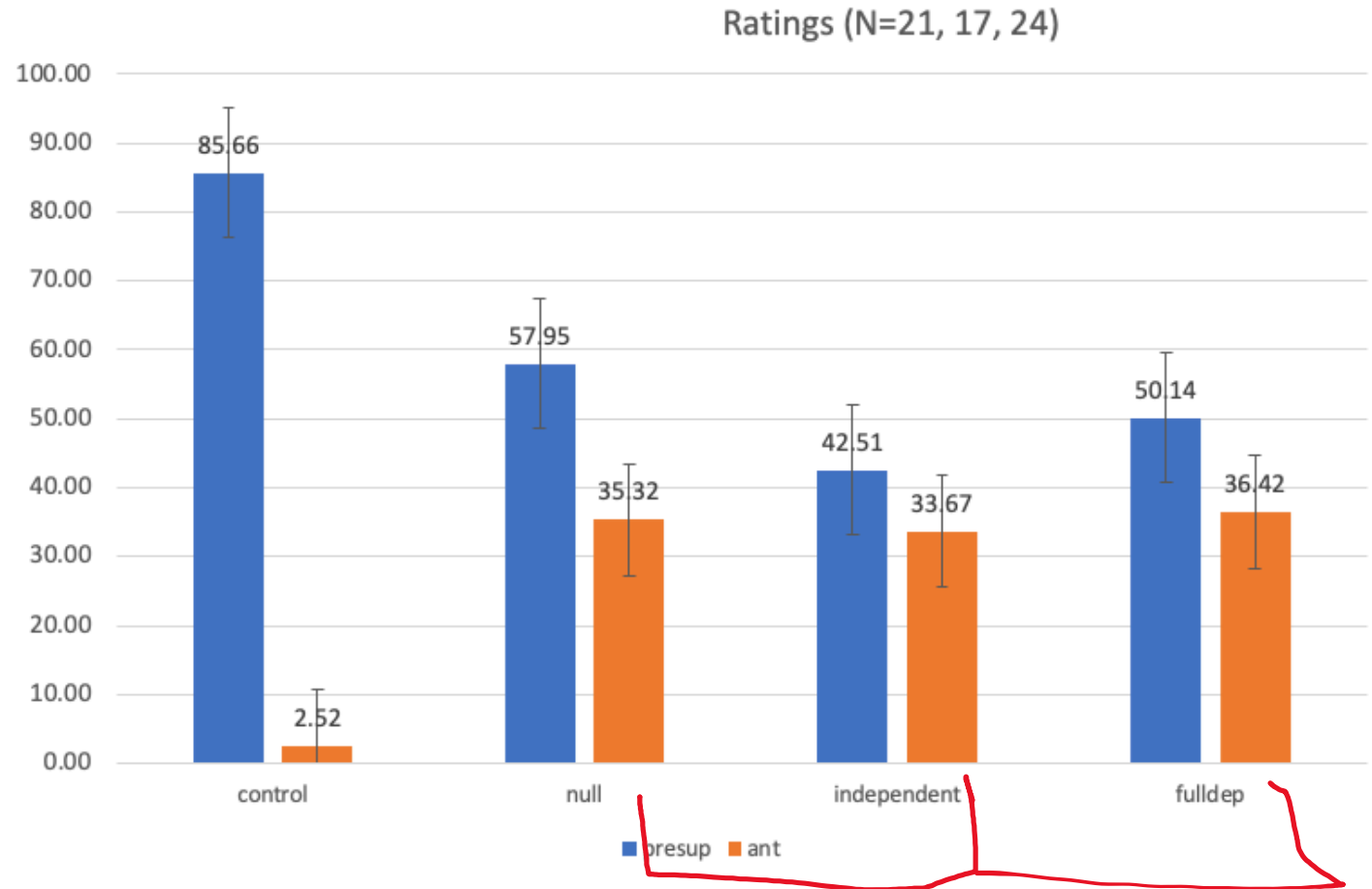 - Breakdown: 21 null, 17 Independent, 24 Fulldep

Ratings: Control

- "John doesn't have a dord" is close to 85, because it was a contradiction

- "John is a glorp" is close to 0, because it was consistent with prior information
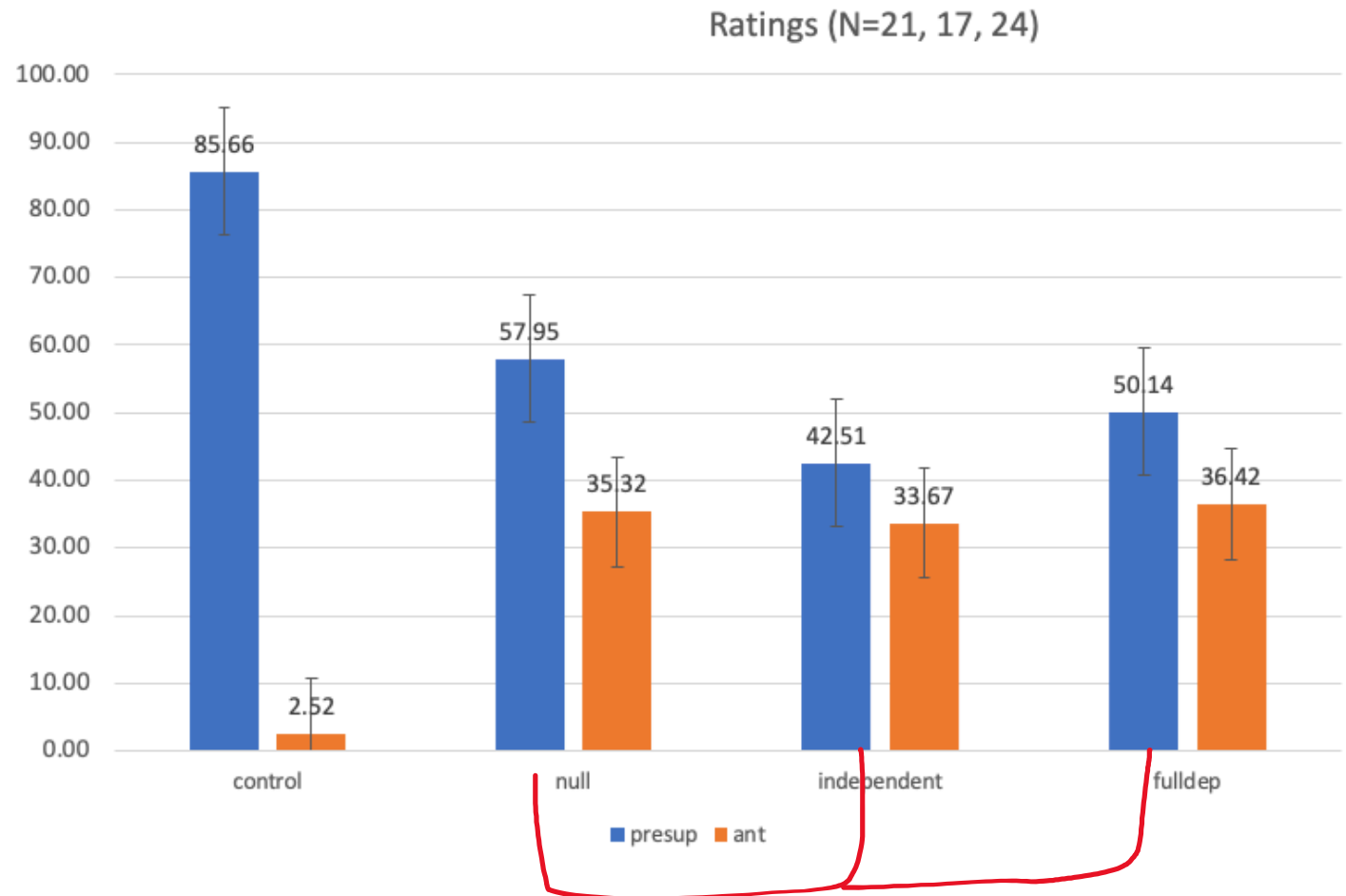
Ratings (N=21, 17, 24)

Ratings: Ant of critical conditions

- Across the 3 versions, "John is a glorp" is around 35

- This is much higher than the **ave** in our pilot 1, which is around 15

- But it's quite similar across the 3 ant conditions
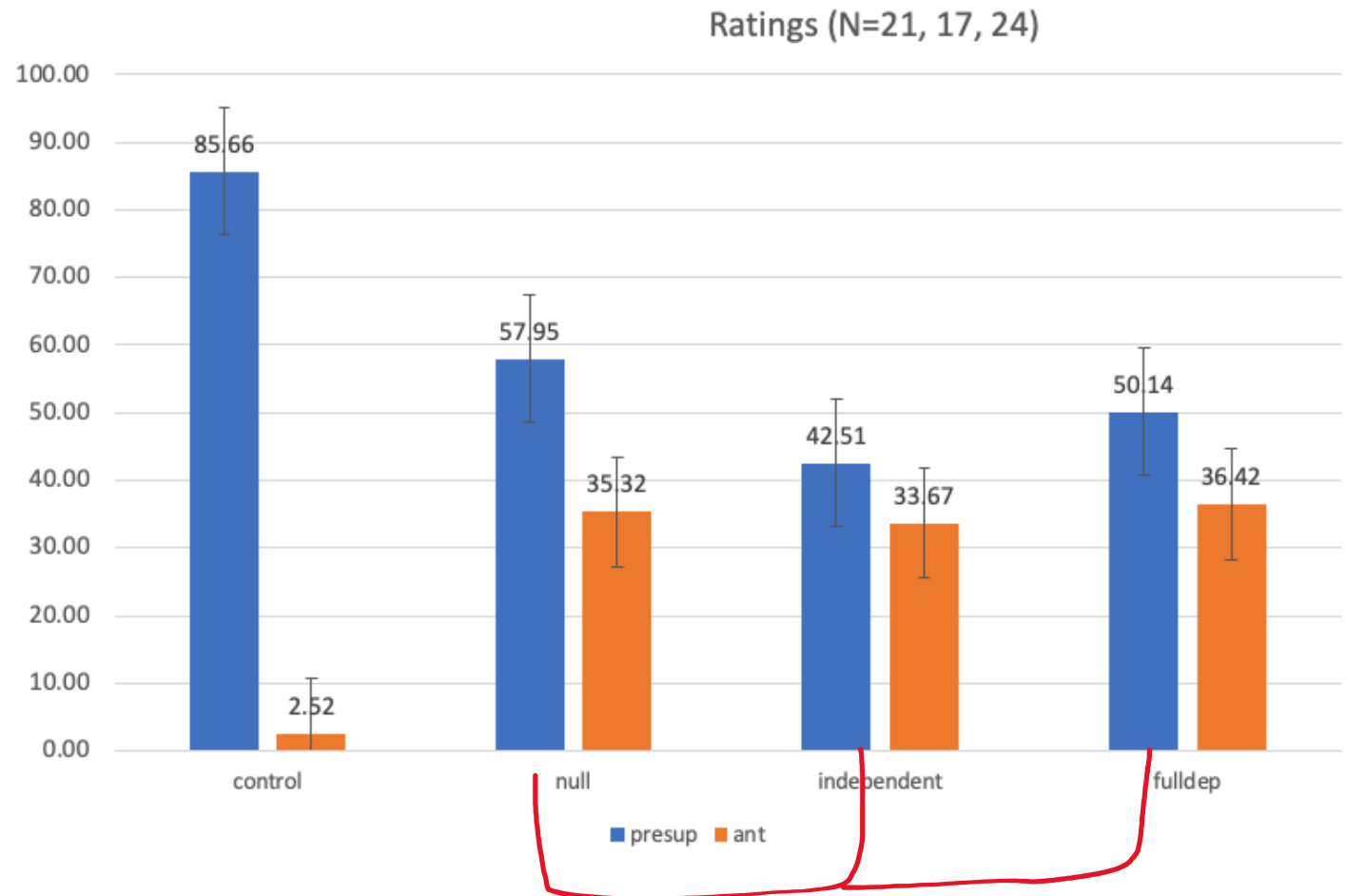

Ratings (N=21, 17, 24)

Ratings: Presup of critical conditions

- **Null.Presup**: 58, the highest ratings of all critical presups; similar to its rating in pilot 2

- **Independent.Presup**: about 42 -- similar to its rating in pilots 1&2

- BUT WEIRDLY: **fulldep** is rated pretty high, and even higher than **independent**???
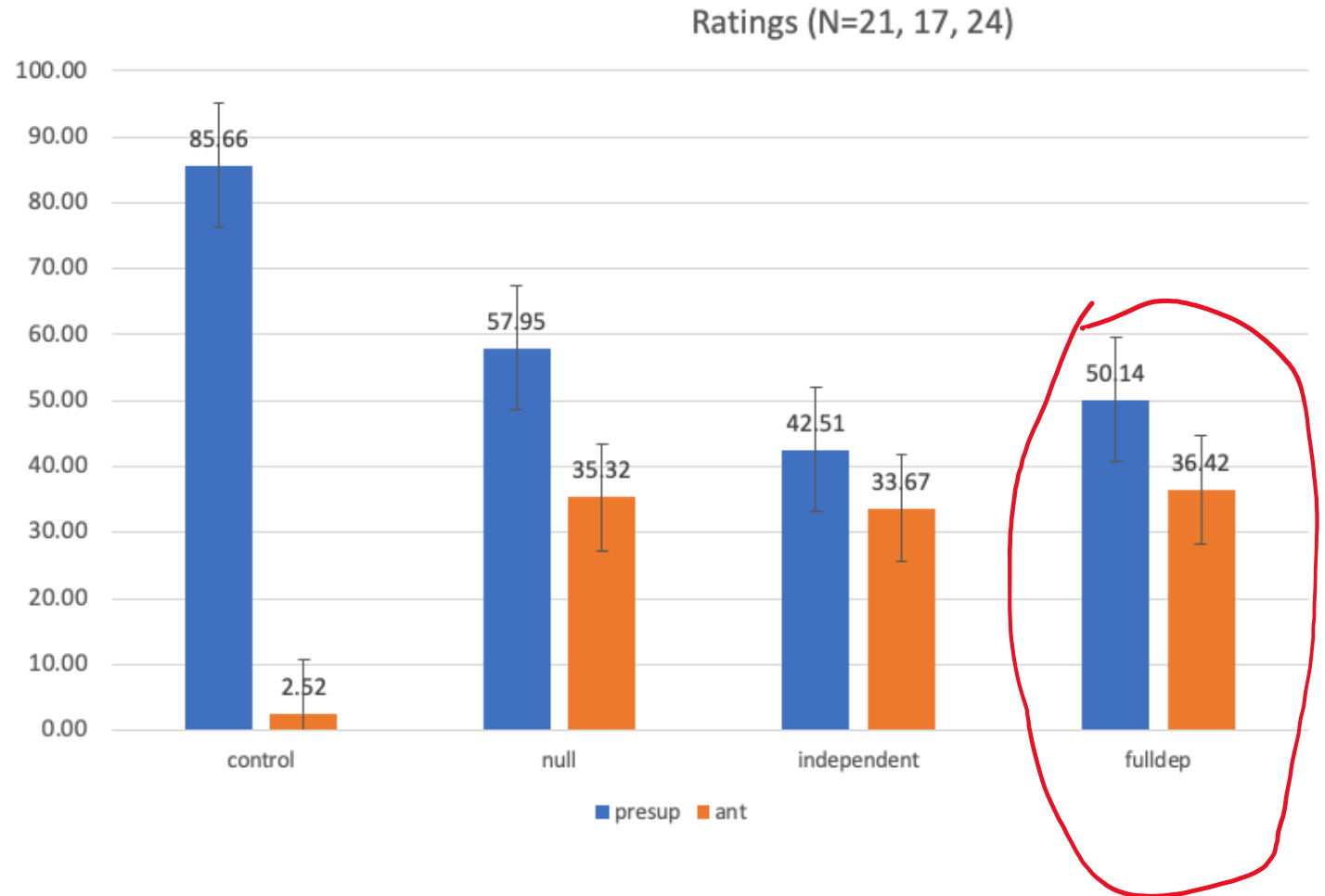
Ratings (N=21, 17, 24)

Ratings: Presup of critical conditions

- While the diff between **null** and **indepdnent** may be interesting (and their respective ratings consistent with previous pilots), it is not really interpretable because of the high ratings of **fulldep**

- If anything, the opposite of what we expected between independent and fulldep

Ratings (N=21, 17, 24)

	presup	ant
control	85.66	2.52
null	57.95	35.32
independent	42.51	33.67
fulldep	50.14	36.42

Ratings: Fulldep

- We expect presup and ant in FullDep to be rated similarly: *if all gogopos have a dord and no one else has dords*, and *"If Cece is a gogopo, she will yapple her dord"*, then whether Cece has a dord = whether Cece is a gogopo

- That was what happened in BOTH pilots 1 & 2. In Pilot 1, both were rated at around 25. Something else is going on here...?

Ratings (N=21, 17, 24)

Chart showing ratings for presup (blue) and ant (orange):
- control: presup 85.66, ant 2.52
- null: presup 57.95, ant 35.32
- independent: presup 42.51, ant 33.67
- fulldep: presup 50.14, ant 36.42

Reflections

- What do we gain & lose from the between-subj design?
  - We don't need to worry about order effects, less likely to be influenced by participants strategizing as they try to figure out what the game is about
  - Meanwhile tho, we also may run the risk of just getting participants who have different intuitions about how to respond to a presup projected out of a conditional, with no direct way of comparing how they may respond differently *depending on the context*
- **FullDep.Presup**
  - It seems to be behaving oddly in this pilot, quite unlike before
  - One might say: since **Null.Presup** and **Independent.Presup** both got ratings consistent with how they were rated when they showed up early (in pilots 1 & 2), this is just how FullDep.Presup looks when it shows up right after the control. We just didn't observe this previously, because it was obscured by some order effects.
  - But its high ratings here still quite unintuitive to make sense of

Reflections

- A compromise that I suggested:
  - Control + null + dep
  - Control + indep + dep (basically pilot 1)

- Why
  - Control for potential order effects to some extent, esp between null+indep
  - But also allows us to see how a participant would respond differently depending on the context